7.1: Learning about the World through Surveys

Definition.
A **population** is a group of objects or people we wish to study.

- A **parameter** is a numerical value describing some aspect of the population (e.g. means and proportions)
- A **census** is a survey of *every member* of a population

A **sample** is a subset of the population of interest.

- A **statistic** is a numerical value describing some aspect of the sample
- Statistics are sometimes called **estimators**
- A **statistical inference** is the science of drawing conclusions about a population based on observing only a small subset of that population.

Example. In February 2014, the Pew Research Center surveyed 1428 cell phone users in the United States who were married or in a committed partnership. The survey found that 25% of cell phone owners felt that their spouse or partner was distracted by their cell phone when they were together.

Identify the population and the sample

Population: everyone with a cell phone

Population: All cell phone users in US, married...

Sample: The 1428 cell phone users in US, married...

Sample: 1428 cell phone users

Identify the parameter and the statistic

Parameter: proportion of cell phone owners from the pop who feel their spouse/partner are distracted

Parameter: Proportion of the population who felt partner was distracted

Statistic: Proportion of sample who felt partner was distracted

Statistic: proportion of cell phone owners from the sample who feel their spouse/partner are distracted

Sample statistics and population parameters are represented using different symbols (English for statistics, Greek for population):

	Statistics	Parameters	
mean	$\overline{x}$	μ	mu
standard deviation	s	σ	sigma
proportion	$\hat{p}$	p	

Example. The City of Los Angeles provides an open data set of response times for emergency vehicles. Each row of the data set represents an emergency vehicle that has been sent to a particular emergency. A random sample of 1000 of these rows shows that the mean response time was 8.25 minutes. In addition, the proportion of vehicles that were ambulances was 0.328.

Using correct notation, identify the data given above.

$n = 1000$

$\bar{x} = 8.25$

$\hat{p} = 0.328$

What can we conclude about the overall population?

The population mean response time will be CLOSE to 8.25
The population proportion of ambulances will be CLOSE to 0.328

Definition.
A method is **biased** if it tends to produce the wrong value.

- **Sampling bias** results from a sample that is not representative of the population.
- **Measurement bias** results from questions that do not produce a true answer.

Example. For the following scenarios, identify any bias:

Online reviews (Amazon, Yelp, etc.)

Sampling bias

Sampling bias

Asking if people support a ‘fat tax’ on non-diet sugary soft drinks

Measurement bias

Measurement bias

Gallup poll calling landline phones

Sampling bias

Sampling bias

Using a poorly written question (e.g. double negative)

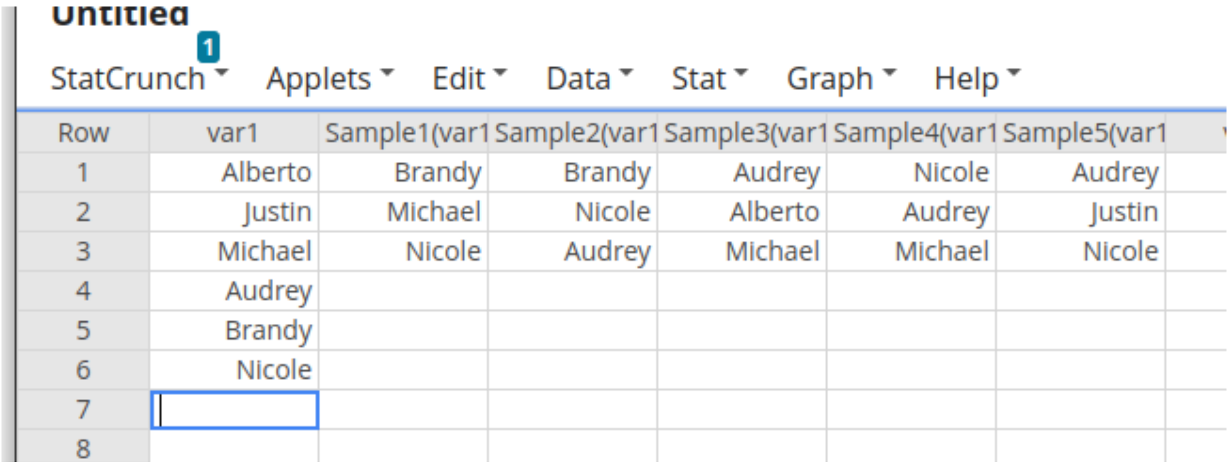
Measurement bias

Measurement bias

Definition.
A **simple random sample (SRS)** is where subjects from a population are drawn *at random* and *without replacement.* With an SRS, each member of the population has an equally likely chance of being selected.
Nonresponse bias results from people refusing to respond to the survey.

Example. Perform an SRS of 3 people from the list below:

1	Alberto
2	Justin
3	Michael
4	Audrey
5	Brandy
6	Nicole



Row	var1	Sample1(var1	Sample2(var1	Sample3(var1	Sample4(var1	Sample5(var1
1	Alberto	Brandy	Brandy	Audrey	Nicole	Audrey
2	Justin	Michael	Nicole	Alberto	Audrey	Justin
3	Michael	Nicole	Audrey	Michael	Michael	Nicole
4	Audrey					
5	Brandy					
6	Nicole					
7						
8						

7.2: Measuring the Quality of a Survey

The true population proportion can be estimated by the sample proportion. How accurate can we expect our estimate to be?

- The *accuracy* of an estimation method is measured in terms of *bias*
- The *precision* of an estimation method is measured in terms of *standard error*

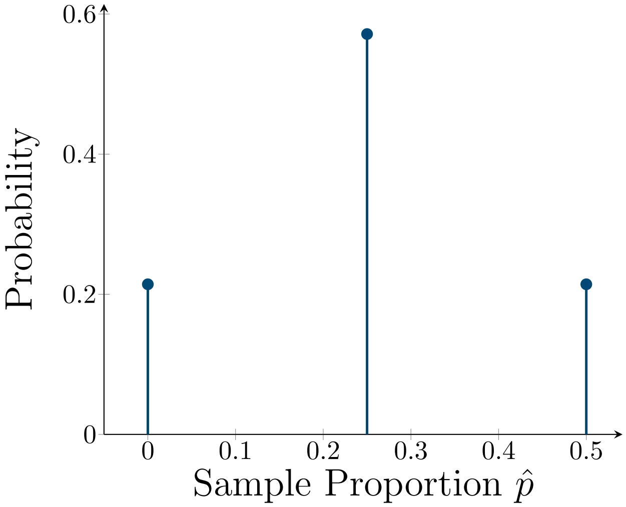
Example. Consider a group of 8 people, where 2 identify as female, and 6 identify as male. What is the true population proportion of females? When using a sample size of $n = 4$, what are possible sample proportions?

FFMMMMMM p=2/8=1/4=0.25

Samples:

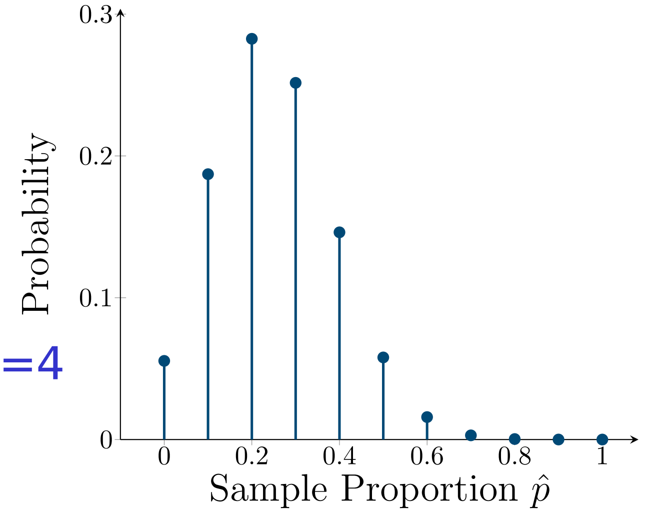
MMMM -> p_hat=0
FMMM, MFMM, ... -> p_hat=0.25
FFMM, MFMF,... -> p_hat=0.5

Accurate? Yes
Precise? No



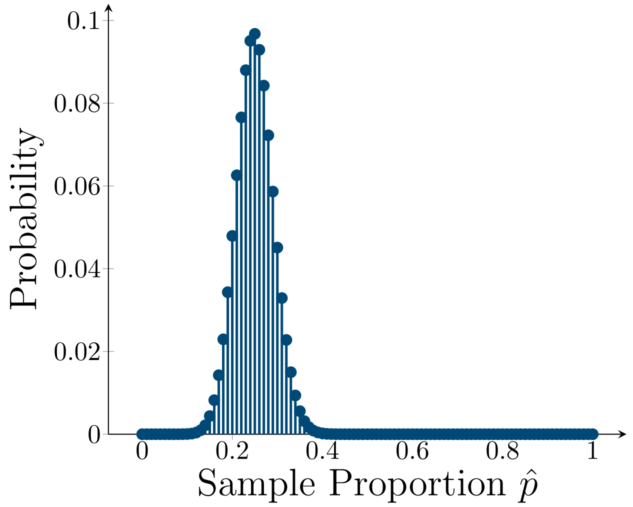
Example. Now, consider a group of 1000 people where 25% identify as female ($p = 0.25$). When using a sample size of $n = 10$, what are possible sample proportions?

Accurate? Yes
Precise? More precise than when n=4



Finally, consider a group of 1000 people where 25% identify as female ($p = 0.25$). When using a sample size of $n = 100$, what are possible sample proportions?

Accurate and
much MORE precise



Definition.

- The **sampling distribution** is the probability distribution of $\hat{p}$.
- The **standard error** for $\hat{p}$ is given by

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

provided that

  - The sample is randomly selected from the population of interest.
  - If sampling without replacement, the population needs to be much larger than the sample size (e.g. at least 10 times bigger)

Since the true population proportion is typically unknown, we can estimate the standard error:

$$SE_{est} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Note: Larger sample sizes have smaller standard error!

7.3: The Central Limit Theorem for Sample Proportions

Definition. (Central Limit Theorem (CLT))

When estimating a population proportion, p, if

1. *Random and Independent*: The sample is collected randomly from the population, and observations are independent of each other.
2. *Large Sample*: The sample size, n, is large enough that the sample can have at least 10 successes or failures.
3. *Big population*: If the sample is collected without replacement, then the population size must be at least 10 times bigger than the sample size.

then the sampling distribution for $\hat{p}$ is approximately Normal, with mean p and standard deviation

$$SE = \sqrt{\frac{p(1-p)}{n}}.$$

This distribution is denoted as

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right).$$

Example. Consider the groups from the previous section where $p = 0.25$ of the group identified as female. Suppose that $\hat{p} = 0.25$. If N represents the population size, and n the sample size, identify if the CLT can be applied.

$N = 8, n = 4$

1. SRS ✓ By assumption

2. $n\hat{p} \geq 10$ ✗ $4(0.25) = 1 \ngeq 10$

   $n(1-\hat{p}) \geq 10$ $4(1-0.25) = 3 \ngeq 10$

Since the 2nd condition is not met, the CLT cannot be applied

3. $N \geq 10n$

$N = 1000, n = 10$

1. SRS ✓ By assumption

2. $n\hat{p} \geq 10$ ✗ $10(0.25) = 2.5 \ngeq 10$

   $n(1-\hat{p}) \geq 10$ $10(1-0.25) = 7.5 \ngeq 10$

Since the 2nd condition is not met, the CLT cannot be applied

3. $N \geq 10n$

$N = 1000, n = 100$

1. SRS ✓ By assumption

2. $n\hat{p} \geq 10$ ✓ $100(0.25) = 25 \geq 10$

   $n(1-\hat{p}) \geq 10$ $100(1-0.25) = 75 \geq 10$

Since all 3 conditions are met, the CLT can be applied

3. $N \geq 10n$ ✓ $1000 \geq 10 \cdot 100$

Example. Consider the group of 1000 people where $p = 0.25$ identified as female. In a sample of $n = 100$ people, what is the probability that $\hat{p}$ is at least 29%?

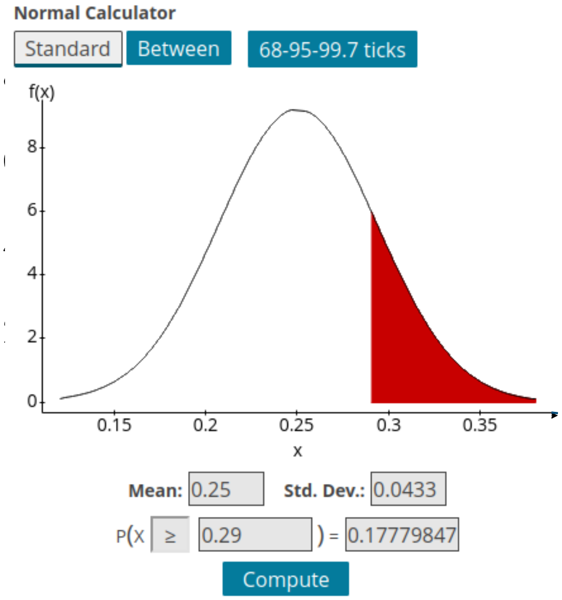
1. SRS ✓

2. $n\hat{p} \ge 10$ ✓ $100(0.29) = 29 \ge 10$
$n(1-\hat{p}) \ge 10$ $100(1-0.29) = 61 \ge 10$

3. $N \ge 10n$ ✓ $1000 \ge 10 \cdot 100$

CLT → $N\left(\underbrace{p}_{0.25}, \underbrace{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}_{0.0433}\right)$

$$\boxed{P(\hat{p} \ge 0.29) = 0.1778}$$


Normal Calculator

Mean: 0.25 Std. Dev.: 0.0433

P(X ≥ 0.29) = 0.17779847

Example. Samuel Morse claimed that the true proportion of E's used in the English language is 0.12. Suppose we take a sample of 876 letters, and find a sample proportion of 0.1347. If we took another sample, what is the probability that the new sample proportion would be greater than 0.1347?

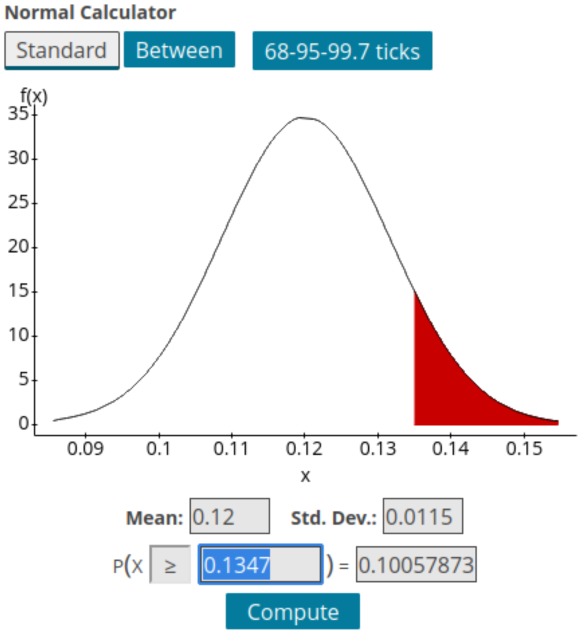
1. SRS ✓

2. $n\hat{p} \ge 10$ ✓ $876(0.1347) = 117.9972 \ge 10$
$n(1-\hat{p}) \ge 10$ $876(1-0.1347) = 758.0028 \ge 10$

3. $N \ge 10n$ ✓ $N \ge 10(876)$

CLT → $N\left(\underbrace{p}_{0.12}, \underbrace{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}_{0.0115}\right)$

$$\boxed{P(\hat{p} \ge 0.1347) = 0.1006}$$


Normal Calculator

Mean: 0.12 Std. Dev.: 0.0115

P(X ≥ 0.1347) = 0.10057873